ABSTRACT

A method for recognizing a table structure from delineated table region in

5   an electronic document using hierarchical clustering of data strings. The cluster

groupings are segregated effectively using the distances from a positional vector

associated with words and groups of words rather than a minimum number of

blank spaces between words. Once a data tree of the hierarchical clusterings is

constructed, the tree is scanned downward from the root to find appropriate

10   column boundaries using a columnization algorithm. Then using successive

heuristic algorithms, determine column and row headers and row boundaries.